The effects of added salt on the second virial coefficients of the complete proteome of *E. coli*

Richard P. Sear[1]

[1] *Department of Physics, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom*

email: r.sear@surrey.ac.uk

Bacteria typically have a few thousand different proteins. The number of proteins with a given charge is a roughly Gaussian function of charge — centered near zero, and with a width around ten (in units of the charge on the proton). We have used the charges on E. coli's proteins to estimate the changes in the second virial coefficients of all its proteins as the concentration of a 1:1 salt is increased. The second virial coefficient has dimensions of volume and we find that on average it decreases by about twice the average volume of a protein when the salt concentration is increased from 0.2 to 1 Molar. The standard deviation of the decrease is of the same order. The consequences of this for the complex mixture of proteins inside an E. coli cell, are briefly discussed.

I. INTRODUCTION

The genomes of a number of organisms are already known and more are being completed at a rate of perhaps one a month. Once a genome[1] has been sequenced the amino-acid sequences of *all* its proteins are known. The complete set of proteins of an organism is generally referred to as its proteome. Here we use genome data for *E. coli* to systematically estimate the change in the interactions of all the proteins of this bacterium when the salt concentration is varied. *E. coli* can grow in environments with a very wide range of salt concentrations[2], and so its proteins must function *in vivo* over a wide range of salt concentrations. The potassium ion concentration inside the cell can vary from approximately 150 to 300 mM[2]; potassium is the predominant cation in living cells. Clearly, the proteins must remain soluble over this range, and they should bind to the other proteins which they are required to bind to in order to function, but they should not interact strongly with other proteins. The study of a proteome is often called proteomics. Here, as the physical properties of proteins are studied (as opposed to their chemical properties such as catalytic function) we are doing what may be called physical proteomics.

There has been extensive theoretical work on the salt dependence of the interactions in individual proteins, particularly for the protein lysozyme[3,4,5]. See Refs. 3,4,6,7,8,9 for corresponding experimental work. However, as far as the author is aware, this is the first attempt to characterise the interactions of *all* the proteins of an organism. We will consider the proteins separately, i.e., as single component solutions. Of course inside a bacterium the proteins exist as a mixture of thousands of components. Future work will consider multi-component mixtures of the proteins. We have chosen *E. coli* as it is a bacterium, and therefore a relatively simple organism, and as it has been extensively studied. However, the distribution of charges on the proteins of almost all organisms is very similar and so our results apply to almost all organisms, including *H. sapiens*. The only exceptions are some extremophiles[10].

In the next section we use genome data to estimate the charges on the proteins of *E. coli*. This data is used in the third section where we calculate the variation in their second virial coefficients as the salt concentration is varied. The last section is a conclusion.

II. THE CHARGES ON PROTEINS OF E. COLI

E. coli K-12 has a proteome of 4358 proteins. The amino acid sequences of all of them are known from the sequencing of its genome[11,12]. *K-12* is the name of a strain of *E. coli*. Runcong and Mitaku have analysed the charge distributions of a number of organisms using a simple approximate method of estimating the charge on a protein at neutral pH from its amino-acid sequence[10]. We will follow their analysis but use a slightly different approximation for the charge on a protein with a given amino-acid sequence.

Of the 20 amino acids, 5 have pK values such that they should be at least partially charged at neutral pH[13,14]. These are two highly acidic amino acids, aspartic acid and glutamic acid, two highly basic amino acids, lysine and arginine and one somewhat basic amino acid, histidine. Aspartic and glutamic acids have pK's far below 7 and lysine and arginine have pK's far above 7 and so we assume that all 4 of these amino acids are fully charged at neutral

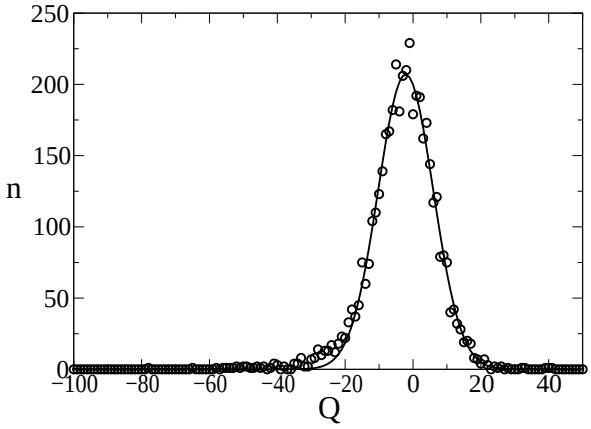

FIG. 1: The number of *E. coli K-12* proteins n as a function of net charge Q. There are 2 proteins with net charges < -100; these are not shown. The data for each value of Q are shown as circles and the curve is a Gaussian fitted to the data.

pH. Aspartic and glutamic acids then each contribute -1 to the charge on a protein, and lysine and arginine each contribute $+1$. Histidine has a pK of around 6-6.5 (this will depend on the environment of the amino acid). The equation for the fraction f of a basic group such as histidine that is charged at a given pH is

$$f = 1/\left(1 + 10^{pH-pK}\right), \tag{1}$$

where pK is the pK value for the basic group. This equation is just the Henderson-Hasselbalch equation[14] rearranged. Taking pK= 6.5[14] and at pH=7 we have that the fraction of histidines charged is $f = 0.24$. As this is small we assume for simplicity that all the histidine amino acids are uncharged.

Thus, with these assumptions for the charges on these 5 amino acids, our estimate for the net charge Q on a protein is simply given by

$$Q = n_K + n_R - n_D - n_E \tag{2}$$

where n_K, n_R, n_D and n_E are the the protein's total numbers of lysines, arginines, aspartic acids and glutamic acids, respectively. The subscripts K, R etc. correspond to the standard single letter codes for the amino acids[13,14]. The charge Q is in units of e where e is the elementary charge.

Note that Runcong and Mitaku[10] assume that the histidine amino acids contribute $+1$ to the charge, that is the only difference between our analysis and that of Runcong and Mitaku. As the histidine amino acid is quite a rare amino acid, approximately 1 in 50 amino acids is a histidine, the difference between the results we obtain and those of Runcong and Mitaku[10] is not large but our charges are shifted to more negative values. Using Runcong and Mitaku's approximation the mean charge on a protein is 7.11 units more positive than the mean charge we find here. As a check on our algorithm, we can compare the prediction of equation (2) for chicken lysozyme to that of a titration experiment to determine the charge. Equation (2) predicts that chicken lysozyme[15] has a net charge of 8 at neutral pH. Titration experiments on lysozyme give a titratable charge of close to 8.5 at pH=7[16].

Using equation (2) we can obtain estimates for the charges of all 4358 proteins of *E. coli*[11,12,17,18]. The results are shown in Fig. 1, where we have plotted the number of proteins n as a function of net charge Q. The distribution is centered almost at a net charge $Q = 0$, and for not-too-large $|Q|$ the distribution is roughly symmetric and Gaussian. The mean charge is -3.15. Given the approximate nature of our equation for the charge on a protein, equation (2), the data is probably consistent with a mean charge of 0. The approximation scheme of Mitaku and Runcong[10] yields a mean charge of $+3.96$. Also, although when $|Q|$ is not too large the distribution can be seen to be reasonably symmetric, *E. coli* has 12 proteins with charges < -50 but none with charges $> +50$. Excluding proteins with very large charges, $|Q| > 30$, the root mean square charge equals 9.1.

A number of other organisms, both other bacteria and eukaryotes such as yeast, have had the charge distribution on their proteomes determined by Runcong and Mitaku[10] and by the author[19]. Almost all of them have a roughly Gaussian distribution centered approximately at zero, like the distribution in Fig. 1. The exceptions are some extremophiles. Extremophiles are organisms that live in extreme environments, for example *Halobacterium sp.* lives in environments with very high levels of salt[20]. The cytosol of *Halobacterium sp.* contains much higher levels of potassium ions than do other organisms so perhaps it is not a surprise that the distribution of charges on its proteins is different.

We have fitted the Gaussian function

$$n(Q) = \frac{1739}{\sigma} \exp(-(Q - \overline{Q})^2/2\sigma^2). \tag{3}$$

to the data for the number of proteins as a function of their charge. It is drawn as the solid curve in Fig. 1. The fit parameters are mean charge $\overline{Q} = -2.16$ and standard deviation $\sigma = 8.32$. 1739 is $4358/(2\pi)^{1/2}$ and so the distribution is normalised so that its integral gives the total number of proteins. Within a couple of standard deviations of the mean the Gaussian function fits the data well but it underestimates the numbers of proteins with charges such that $|Q - \overline{Q}|$ is several times the standard deviation.

We also note that there is a correlation between the net charge Q on a protein and its size, measured by the number of amino acids M. Figure 2 is a scatter plot of charge Q and number of amino acids M for the proteins of *E. coli*. Although at any particular size M there is a wide distribution of charges, on average the more highly charged proteins are larger than average. We expect the volume of a protein to scale with M.

III. SALT DEPENDENCE OF THE SECOND VIRIAL COEFFICIENTS

Consider a dilute solution of a single one of the proteins of *E. coli*. Apart from water, the only other constituents are a 1:1 salt at a concentration c_s and a buffer which controls the pH while making a negligible contribution to the ionic strength. Here we will always assume the pH=7 but other pH's can be considered if the net charges on the proteins can be calculated. Also, the counterions of the protein are assumed the same as either the anions or cations of the salt, depending on the sign of Q. The interactions between the protein molecules in the salt solution can be characterised by means of the protein's second virial coefficient B_2: a function of temperature, pH and salt concentration.

Proteins are complex molecules and we are unable to calculate from first principles the absolute value of B_2 for any of the 4358 proteins possessed by *E. coli*. However, predicting the change in the second virial coefficient when the salt concentration varies is a much easier problem, *if* we assume that changing the salt concentration changes only the direct electrostatic interaction between the net charges of a protein. This is a strong assumption but studies of the simple protein lysozyme have shown that the variation of its second virial coefficient can be described using a simple model which only includes its net charge[3,5]. Here we will follow Warren[5] and apply his analysis of lysozyme to the complete set of proteins of *E. coli*. We will discuss which proteins are likely to be less well described by this theory than is lysozyme.

A protein molecule of charge Q is surrounded by its counterions and as the concentration of the protein is increased so is the counterion density. This increase in the counterion density decreases the translational entropy of the counterions and this contributes a positive amount to the second virial coefficient. See Warren[5] and references therein for details. B_2 has the form[5]

$$B_2 = B_2^{(ne)} + \frac{Q^2}{4c_s}, \tag{4}$$

where $B_2^{(ne)}$ is an assumed constant term due to excluded volume interactions and other interactions which are insensitive to salt concentration. The second term is from the counterions and the salt. It is quadratic in the charge and so of course is zero for uncharged proteins and is independent of the sign of the net charge on a protein.

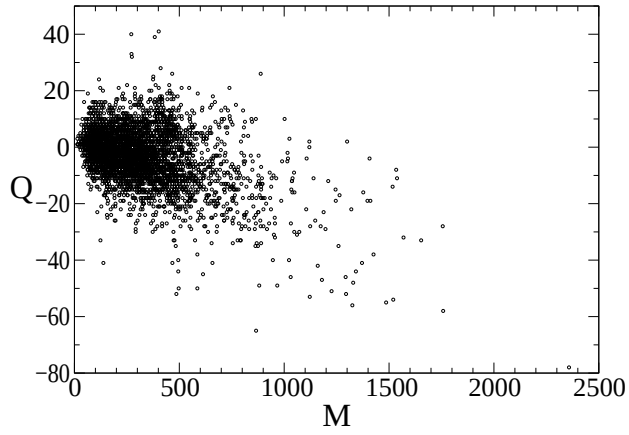

FIG. 2: A scatter plot of the charge on a protein Q versus the number of amino acids M. All but 2, both with charges < -100, of the proteins of *E. coli K-12* are shown.

As stated above we are unable to calculate the absolute value of B_2; the values of $B_2^{(ne)}$ of the proteins are unknown. However, we can calculate the difference in B_2 when the salt concentration is changed from $c_s = c_1$ to $c_s = c_2$. It is

$$\Delta B_2 = \frac{Q^2}{4}\left(\frac{1}{c_2} - \frac{1}{c_1}\right). \tag{5}$$

This is easy to calculate for any protein and in Fig. 3 we have plotted the number of proteins n as a function of the change in their second virial coefficient, ΔB_2, when the salt concentration is decreased from 1M to 0.2M. The results are given in units of nm^3. For comparison the volume of a typical bacterial protein is about 60nm^3 and so if a protein were to interact solely via a hard repulsion it would have a second virial coefficient of about 4 times its volume or about 240nm^3.

Results for proteins with $-30 \le Q \le 30$ are shown. Proteins with larger titratable charges are likely to have an effective charge lower than Q, see Refs. 21,22 and references therein. From the linear Poisson-Boltzmann equation, the potential (divided by e) at the surface of a spherical particle with charge Q and radius a is $Q\lambda_B kT/((1+\kappa a)a)$. $\lambda_B = e^2/(4\pi\epsilon kT)$ is the Bjerrum length, and and κ^{-1} is the Debye screening length, given by $\kappa^2 = 8\pi\lambda_B c_s$. For the dielectric constant of water 80 times that in vacuum and at room temperature, $\lambda_B = 0.7$nm. Globular proteins are approximately spherical and typically have radii around 2 to 4nm. Taking a protein with a radius of 3nm, in salt at a concentration $c_s = 0.1$M, we have that for $Q = 30$, the potential at the surface is about $2kT$. Larger charges correspond to larger surface potentials and these large potentials bind oppositely charged ions to the surface reducing the effective charge. On average, this effect will be diminished to a certain extent by the fact that the most highly charged proteins are larger than average. See Fig. 2, where it is clear that the charge and size of a protein are correlated. Recent simulations by Lobaskin *et al.*[21] of spheres with radius 2nm and charge $Q = -60$ in the absence of salt found an effective charge of a little under -20. Thus we restrict ourselves to proteins with charges of magnitude less than or equal to 30. 4300 of the 4358 proteins, or almost 99%, have charges in this range. The mean change in B_2 of these 4300 proteins when the salt concentration is decreased from 1 to 0.2M is 139nm^3 and the standard deviation is 218nm^3.

A couple of caveats. The first is that the effect of salt on protein solutions is known to depend not only on whether the salt is a 1:1 salt, a 1:2 salt etc. but also to the nature of ions, whether it is Mg^{2+} or Ca^{2+} for example[23]. Our generic theory applies only where there are no specific interactions between the salt and the protein. There is good agreement between experiment and theory for lysozyme plus NaCl[3,5] and so we may hope that it applies to NaCl and many proteins but it clearly misses potentially important effects for other salts where there are specific protein-salt interactions. The second is that proteins are not simple charged spheres, for example some have large dipole moments. Dipoles exert net attractions which are screened and hence weakened by added salt. Thus proteins

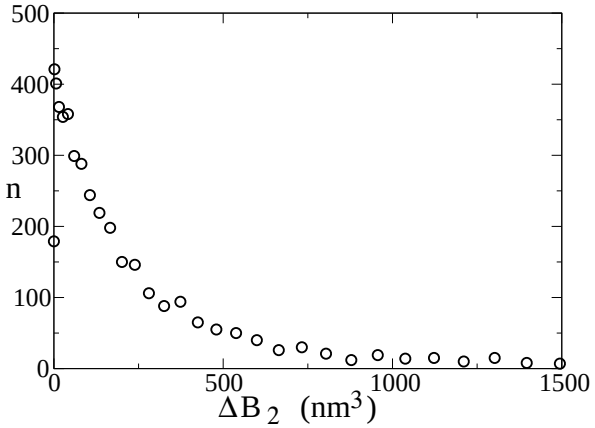

FIG. 3: The number of *E. coli K-12* proteins n as a function of the change in their second virial coefficient, ΔB_2, when the salt concentration is decreased from 1 Molar to 0.2 Molar. Results are only shown for proteins with $|Q| \leq 30$.

with a small charges but large dipole moments are poorly described by the current theory: if the dipole interactions are dominant then the second virial coefficient may even increase when the salt concentration is increased. Velev *et al.*[6] discuss this point. Note that although we can estimate the charge on a protein from its amino-acid sequence we cannot estimate its dipole moment without knowing its three-dimensional structure, and so the sequence data from genomics is not adequate to determine dipole moments.

IV. CONCLUSION

Here we have shown how data from genomics can be used to estimate the charges on the proteins of an organism. We then used these charges to estimate the changes in the second virial coefficients of 4300 (99%) of the proteins of *E. coli* when the salt concentration is changed. Note that *E. coli* can survive and multiply in external environments with a very wide range of salt concentrations; Cayley *et al.*[2] studied the growth of *E. coli* in environments with salt concentrations ranging from very low to 0.5 Molar, corresponding to potassium ion concentrations inside the cell of 150 to 300 mM[2]. Thus, studying the change in interactions of proteins with salt concentration is of direct relevance to the *in vivo* behaviour of proteins. Within molecular biology there is a clear shift of emphasis away from studying the proteins of an organism one or a few at a time, and towards determining the structure and function of large sets of proteins, in particular proteomes. The systematic study of these large sets of proteins is often called proteomics. This work is a first attempt to keep up with this shift by performing a simple theoretical calculation of a solution phase physical property for a complete proteome, rather than for one or a handful of proteins as is usually done. It may be termed physical proteomics.

Future work could consider mixtures of proteins, ultimately aiming to understand the cytosol of a living cell, which is a mixture of of order 10^3 different types of proteins as well as DNA, RNA, ions like ATP and potassium, etc.. This is of course very complex but *if* in the cytosol the proteins of *E. coli* are present in amounts which are uncorrelated with their net charge, the mean charge of the proteins will be close to the mean of the distribution of Fig. 1. This is quite small. Neidhardt[24] has taken an inventory of the species inside *E. coli*, and the results for charged macromolecules are shown in Table I. The charged macromolecules in a cell are protein, DNA and the various forms of RNA: transfer RNA (tRNA), messenger RNA (mRNA) and the RNA in ribosomes (rRNA). See Refs. 13,14 for an introduction to the proteins, DNA and RNA. Although for every molecule of tRNA molecule there are 10 of protein, for every ribosome there are 100 molecules of protein, and for every mRNA there are 1000 molecules of protein, the contributions of the tRNA, ribosomes, DNA and proteins, to the overall charge density of the macromolecules are very roughly comparable. The ribosomes contribute the largest amount. The macromolecules are negatively charged and this negative charge is

Charged species	Q	n	Charge density in cytosol (mM)
Protein	-21 to $+15$	2×10^6	-10
tRNA	-80	2×10^5	-30
mRNA	-2000	1×10^3	-3
Ribosome	-3000	2×10^4	-100
DNA	-10^7	1	-20

TABLE I: The charged macromolecular species in the cytosol of *E. coli*. The data is from Neidhardt[24]. Q is the charge on a macromolecule. For proteins the range given is the mean plus and minus twice the standard deviation. n is the total number of molecules of a species per cell. The volume of an *E. coli* cell is about $10^{-18}\mathrm{m}^3$ so 1 molecule per cell corresponds to a concentration of about 2×10^{-9} Molar. A prokaryote ribosome consists of about 4500 bases of RNA plus protein. The ribosomal proteins are mostly quite stongly positively charged and so will decrease the net negative charge. As a rough estimate we settle on a net charge of -3000. The charge on a tRNA molecule is around -80, -1 from each of its bases[13,14]. The charge on a mRNA molecule is on average around -2000[24]. The charge on DNA is equal to twice the number of base pairs, 4,639,221 for *E. coli K-12*[11]. The charge density from the proteins assumes that the proteins have the mean charge of -3 that they would have if their density and charge were uncorrelated.

balanced by potassium ions[13]. Thus the cytosol resembles a solution of a negatively charged polyelectrolyte, except that there is not one, relatively simple, macromolecular species, but thousands of rather complex and diverse species of macromolecules.

Acknowledgements

It is a pleasure to acknowledge discussions with J. Cuesta, D. Frenkel and P. Warren.

---

[1] A brief introduction to the biological nomenclature: an organism's DNA contains many genes, each of which codes for a protein. The complete set of genes is called the organism's genome and we will refer to the complete set of proteins as its proteome. Some authors use the word proteome somewhat differently, they use it to denote the set of proteins present in the cytosol of an organism at a particular time. See for example[13] for a more detailed definition of a genome.

[2] S. Cayley, B. A. Lewis, H. J. Guttman and M. T. Record Jr, J. Mol. Biol. **222**, 281 (1991).

[3] W. C. K. Poon, S. U. Egelhaaf, P. A. Beales, A. Salonen and L. Sawyer, J. Phys. Cond.: Matt. **12**, L569 (2000).

[4] M. Muschol and F. Rosenberger, J. Chem. Phys. **107**, 1953 (1997).

[5] P. B. Warren, J. Phys.: Cond. Matt. **14**, 7617 (2002).

[6] O. D. Velev, E. W. Kaler and A. M. Lenhoff, Biophys. J. **75**, 2682 (1998).

[7] B. Guo, S. Kao, H. McDonald, A. Asanov, L. L. Combs and W. W. Wilson, J. Cryst. Growth **196**, 424 (1999).

[8] D. F. Rosenbaum, A. Kulkarni, S. Ramakrishnan and C. F. Zukoski, J. Chem. Phys. **111**, 9822 (1999).

[9] R. Piazza and M. Pierno M, J. Phys.: Cond. Matt. **12**, A443 (2000).

[10] K. Runcong and S. Mitaku, Genome Informatics **12**, 364 (2001).

[11] F. R. Blattner *et al.*, Science **277**, 1453 (1997).

[12] The complete proteome of *E. coli K-12*, i.e., the amino-acid sequences of all its proteins, can be downloaded from databases such as that at the European Bioinformatics Institute (http://www.ebi.ac.uk/proteome). *E. coli K-12* was sequenced by Blattner *et al.*[11].

[13] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts and J. D. Watson, Molecular Biology Of The Cell (3rd Edition, Garland Publishing, New York, 1994).

[14] L. Stryer, Biochemistry (4th Edition, Freeman, New York, 1995).

[15] The sequence used to determine this charge has a Protein Data Bank ID[25] of 1GWD.

[16] This value is obtained from Fig. 1 of C. Tanford and R. Roxby, Biochemistry **11**, 2192 (1972).

[17] M. Gerstein and H. Hegyi, FEMS Microbiology Rev. **22**, 277 (1998).

[18] S. Mitaku, M. Ono, T. Hirokawa, S. Boon-Chieng and M. Sonoyama, Biophysical Chem. **82**, 165 (1999).

[19] R. P. Sear, unpublished work.

[20] M. T. Madigan, J. M. Martinko and J. Parker, Brock Biology of Microorganisms (Prentice Hall, New Jersey, 1997).

[21] V. Lobaskin, A. Lyubartsev and P. Linse, Phys. Rev. E **63**, 020401 (2001).

[22] P. B. Warren, J. Chem. Phys. **112**, 4683 (2000).

[23] S. D. Durbin and G. Feher, Ann. Rev. Phys. Chem. **47**, 171 (1996).

[24] F. C. Neidhardt, Chemical composition of *E. coli*, in E. coli *and* S. typhimurium*: Cellular and Molecular Biology* edited by F. .C. Neidhardt J. L. Ingraham, K. B. Low, B. Magasanik and H. E. Umbarger (American Society for Microbiology, Washington D. C., 1987).

[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000). Web site: http://www.rcsb.org/pdb/.